WHAT IS CLAIMED IS:

1. A corpus stored in a computer-readable medium for training a language model, the corpus comprising:

a plurality of characters; and

a plurality of morphological tags associated with a plurality of sequences of characters of the plurality of characters, the plurality of morphological tags indicating a morphological type of an associated sequence of characters and a combination of parts forming a morphological subtype.

2. The corpus of claim 1 wherein the morphological type is one of affixation, reduplication, split, merge and head particle.

3. The corpus of claim 1 wherein the morphological type is an affixation and the combination of parts includes a word and at least one of a prefix and a suffix.

4. The corpus of claim 3 wherein the combination of parts indicates a part of speech for the word.

5. The corpus of claim 1 wherein the morphological type is a reduplication and the

combination of parts includes a pattern of characters.

6. The corpus of claim 1 wherein the morphological type is a merge and the combination of parts includes a pattern of characters.

7. The corpus of claim 1 and further comprising a plurality of factoid tags providing indications of whether a sequence of characters is a factoid.

8. The corpus of claim 1 and further comprising a plurality of named entity tags providing indications of whether a sequence of characters is a named entity.

9. The corpus of claim 1 and further comprising an indication of whether a sequence of characters is contained in a lexicon.

10. A computer readable medium having instructions for performing word segmentation, the instructions comprising:

receiving an input of unsegmented text;

accessing a language model to determine a segmentation of the text;

detecting a morphologically derived word in the text; and

providing an output of segmented text and an indication of a combination of parts that form the morphologically derived word.

11. The computer readable medium of claim 10 wherein the instructions further comprise indicating that the morphologically derived word is one of an affixation, reduplication, split, merge and head particle.

12. The computer readable medium of claim 11 wherein the instructions further comprise detecting a lexicon in the text.

13. The computer readable medium of claim 10 wherein the instructions further comprise detecting a factoid in the text.

14. The computer readable medium of claim 10 wherein the instructions further comprise detecting a named entity in the text.

15. The method of claim 10 wherein providing an output further comprises indicating a part of speech for the combination of parts.

16. The method of claim 10 wherein providing an output further comprises indicating a pattern of characters forming the combination of parts.

17. A method of developing a corpus for training a language model, comprising:

extracting a list of potential words from a corpus that match defined words and rules;

determining if the list includes a sufficient number of defined words and rules;

annotating the corpus to provide indications of word type; and

providing morphological tags in the corpus indicating a morphological type of an associated sequence of characters and a combination of parts forming a morphological subtype.

18. The method of claim 15 wherein annotating further comprises providing indications of whether the word is a lexicon, a morphologically derived word, a factoid and a named entity.

19. The method of claim 17 wherein the morphological type is one of affixation, reduplication split, merge and head particle.

20. The method of claim 17 wherein providing morphological tags further comprises indicating a part of speech for the combination of parts.

21. The method of claim 17 wherein providing morphological tags further comprises indicating a pattern of characters for the combination of parts.

22. The method of claim 17 and further comprising, after providing morphological tags in the corpus, using said corpus to annotate a larger amount of text.